Intro to sequencing

Introduction to systems biology

13th February 2013

Rachita Yadav

Agenda

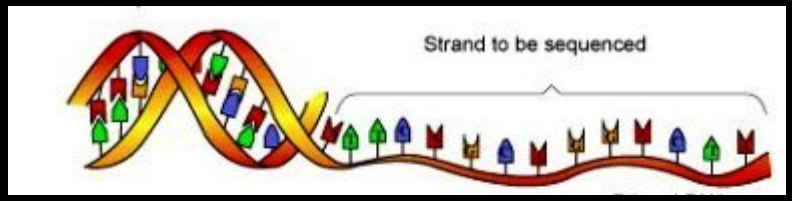
- Sequencing
- Technique
- Data files and formats
- Data analysis
- Introduction to exercise I

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Sequencing

Sequencing

- Sequencing is decoding of the order of bases in DNA or RNA molecules



- Human genome has 3 billion base pairs.

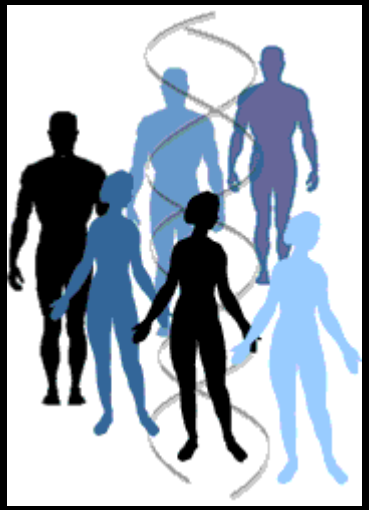
- We are 99.9% identical, but the 0.1% difference is important in affecting our susceptibility to diseases and responsiveness to medication.

AGTCCGCGAATACAGGCTCGGT

- The 0.1% difference comes in many forms such as Single Nucleotide Polymorphisms (SNPs), Insertions, Deletions, Translocations etc.

- Sequencing humans (and a large number of them) will help us develop better therapies for some of the common and life threatening diseases.

- Sequencing animals and plants will enable us to better utilize them to meet the ever growing demand for food and energy.



Types of sequencing

- 1st generation
  - Sequenced individual clones, long continuous reads (>800 nucleotides)
- 2nd generation ~ next generation sequencing
  - Sequence up to tens of thousands of molecules in parallel
  - Shorter read length (~25-700 nucleotides)
- 3rd generation
  - Longer read length
  - Sequencing a human genome in three minutes
  - Price of a human genome for $5,000

Next Generation Sequencing – Applications

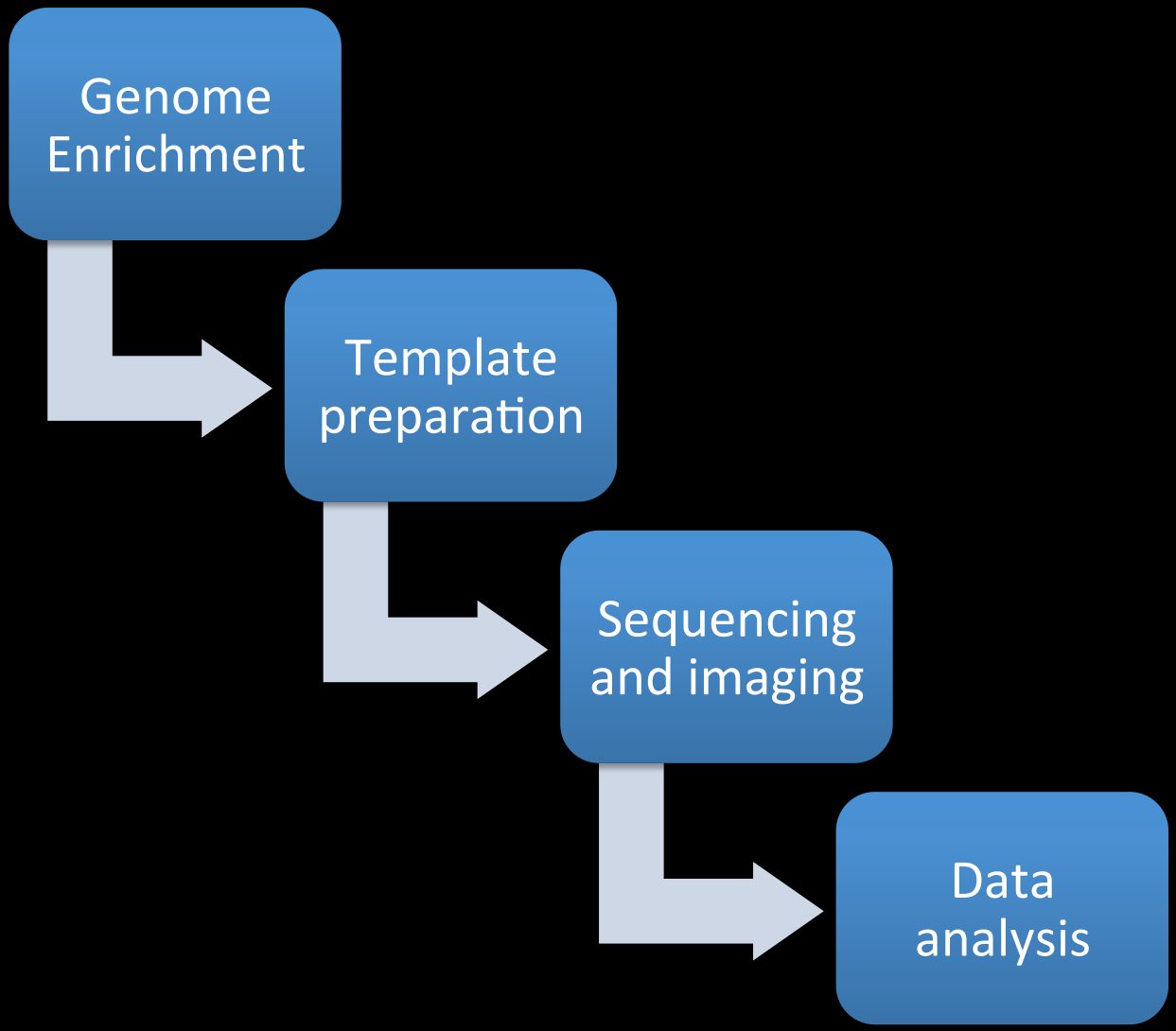
CENTERFO RBIOLOGI CALSEQU ENCEANA LYSIS CBS

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Technique

Steps of next generation sequencing

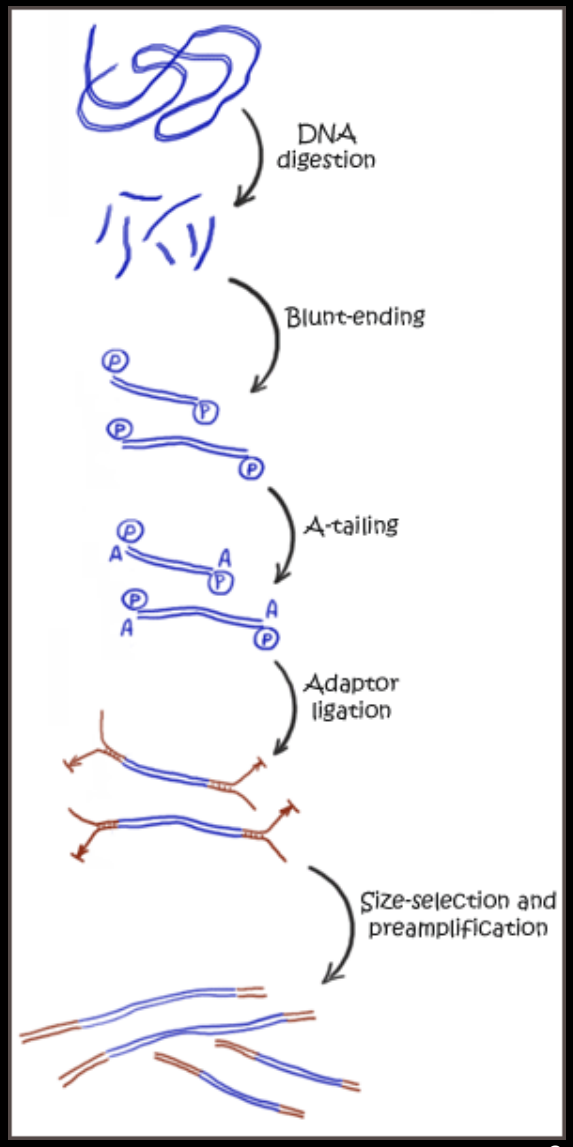
CENTERFOR BIOLOGICAL SEQUENCE ANALYSIS CBS



Technique

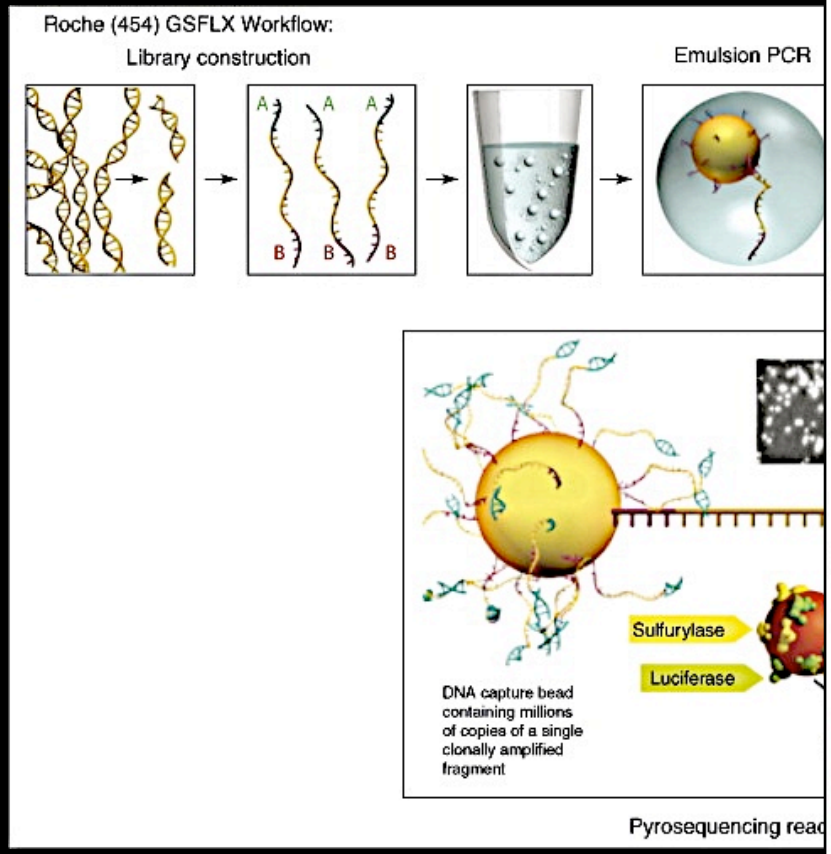
Library preparation

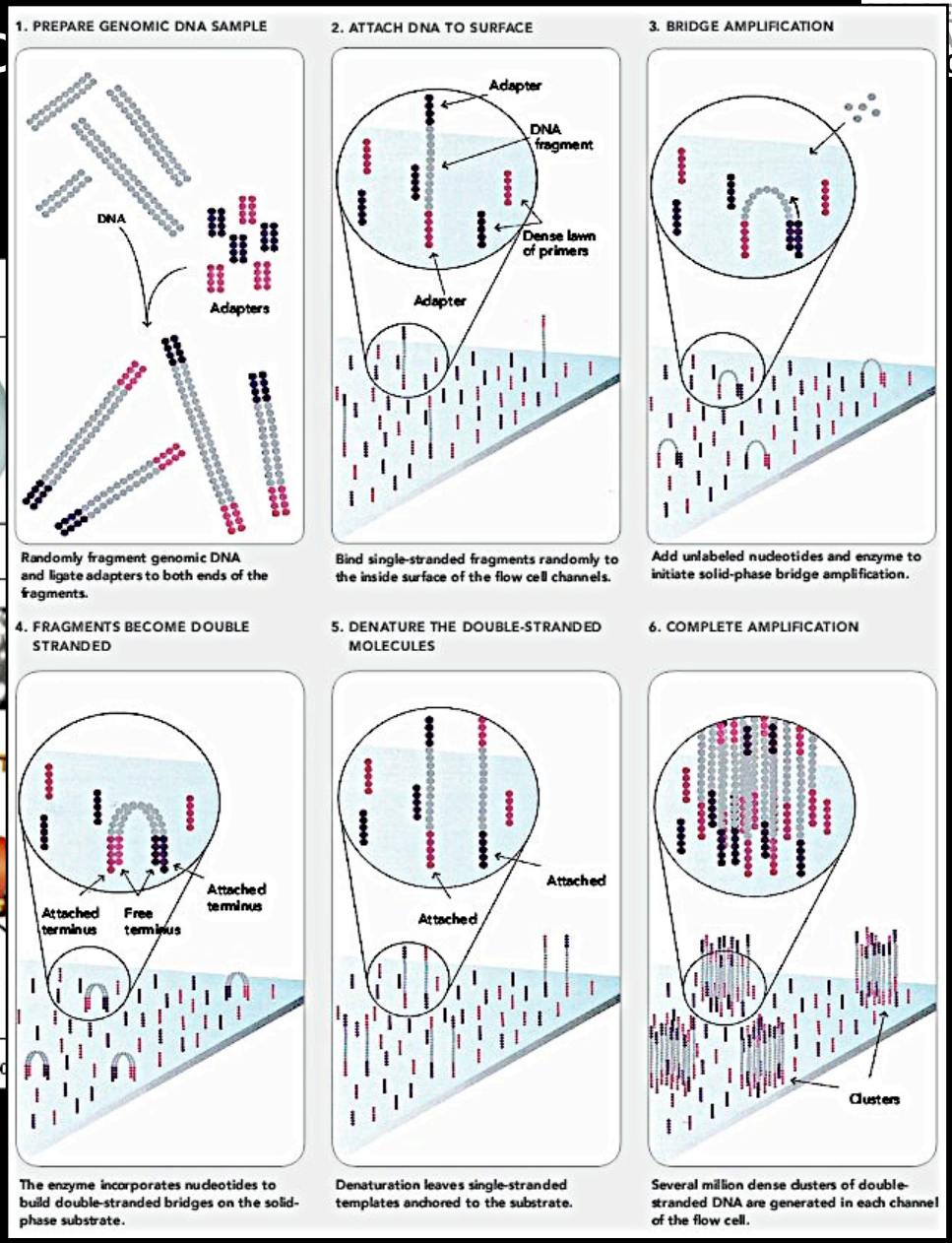
1. Isolate molecule of interest
2. Digest molecule
3. A-tailing: prevents self-ligation and ligation to other fragments
4. Adaptor ligation for PCR amplification and sequencing
5. Amplify library
6. Size-selection



http://seq.molbiol.ru/sch_lib_fr.html

Tec
2nd gen

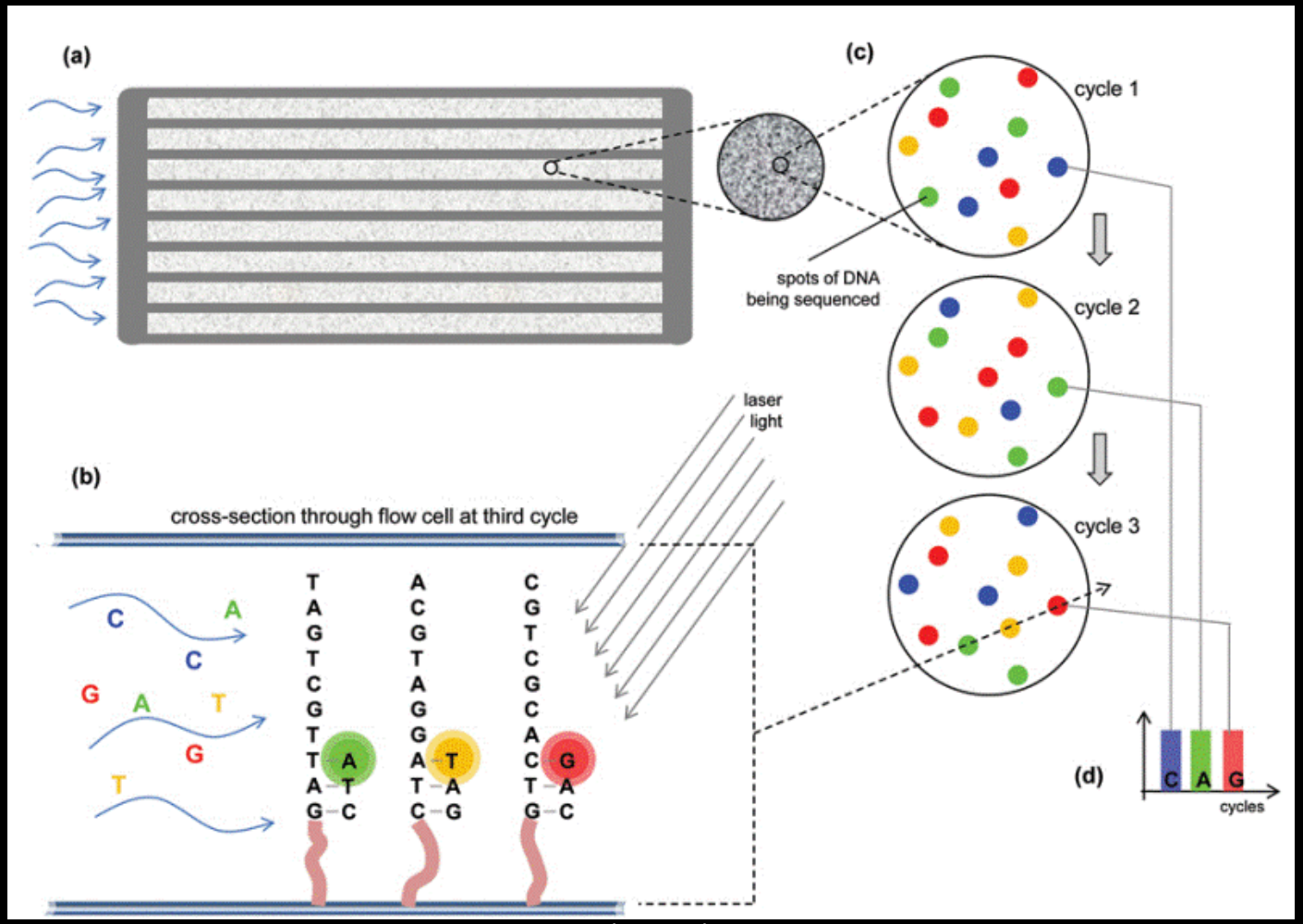




Elaine R. Mardis, Trends in Genetics, 2007

http://www.illumina.com/

Technique
2nd generation seq



2nd generation sequencing

Data files and formats

Read Terminology

- Single end reads – Reads generated from sequencing one end of the fragment of interest

- Paired end reads – Reads generated from sequencing both the ends of the fragment of interest
  - Mate pair library – Generated by circularizing 2-5 kb DNA fragment.

- FASTA/FASTQ – **FASTA** is a standard text format for storing sequence information, in which base pairs are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. FASTQ is similar to FASTA with the addition of quality scores for the bases.

FASTQ format

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

- A FASTQ file normally uses four lines per sequence.
  - Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description.
  - Line 2 is the raw sequence letters.
  - Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
  - Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

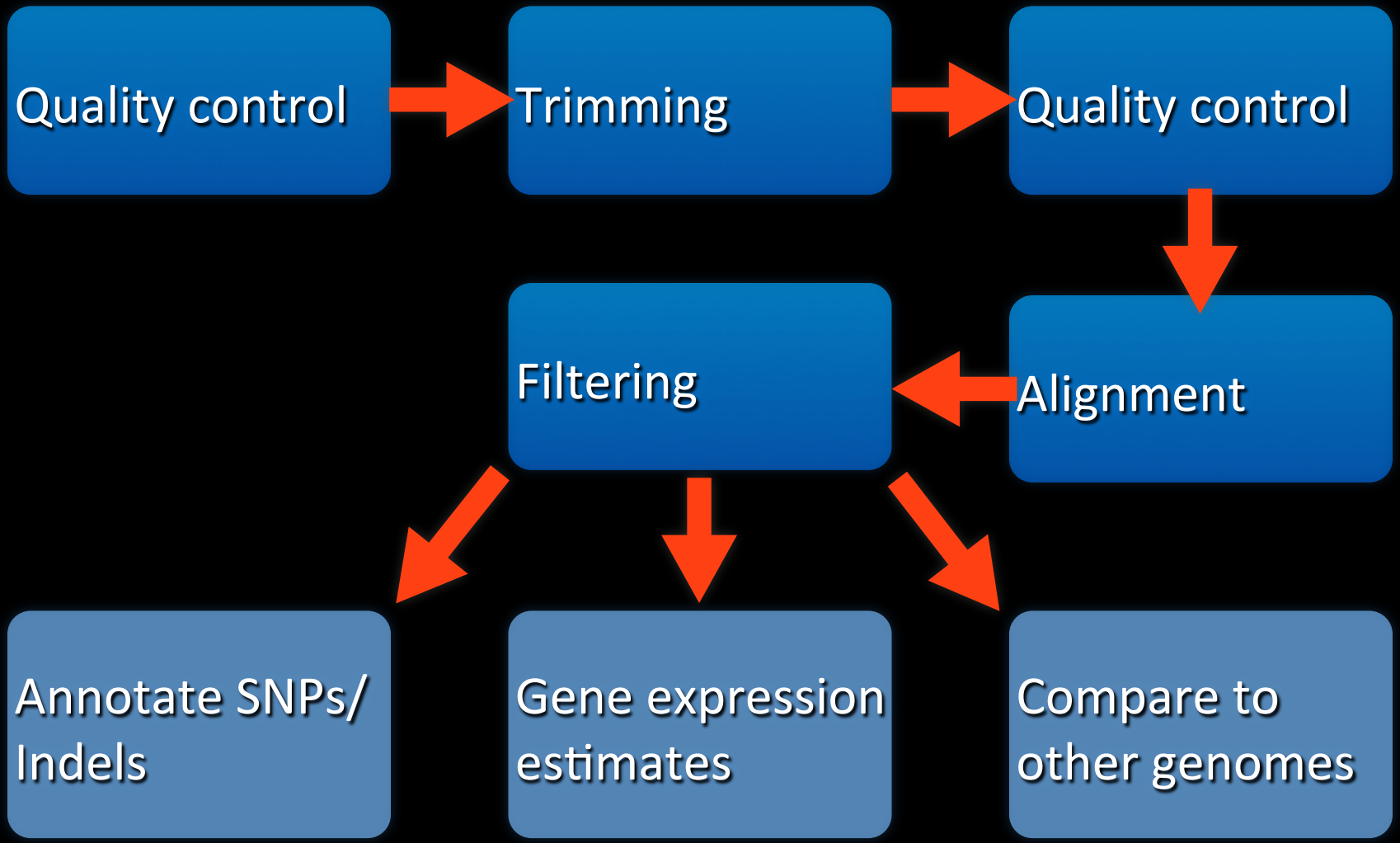
- A minimal FASTQ file might look like this:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

- The quality scores are represented by ASCII characters.
  - In Sanger format, the quality scores are offset by 33 (quality score of 1 to 93 translates to 33 to 126)
  - In Illumina format, depending on the version of pipeline software used, the quality scores are offset by 64, 66 or 33

Data analysis

Workflow of analysis



Files as in system:

Paired-end sequencing

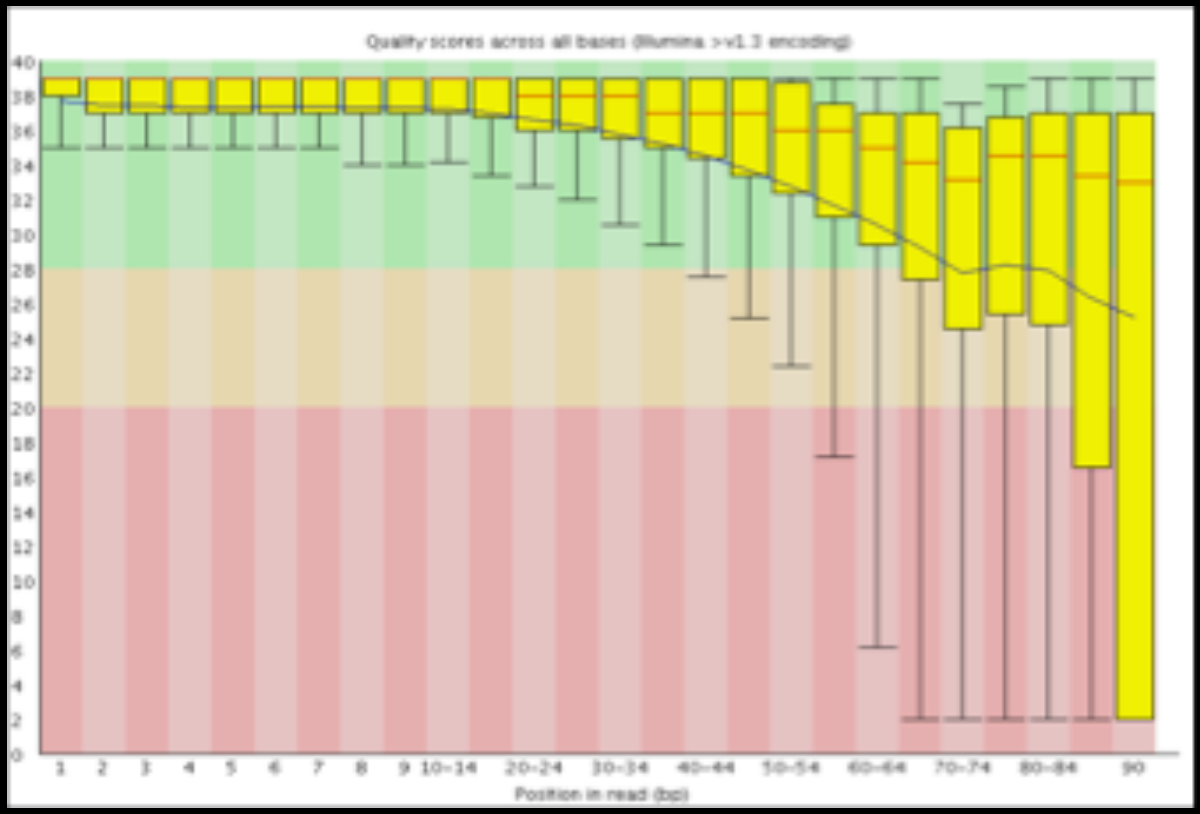
- Each fastq file contains ~95,000,000 lines
- File suffixes (_1 &_2) show they have corresponding reads.

```
interaction[belling]:/home/panfs/cbs/projects/breastcancer/belling/RNA_seq_data> ll
total 46039552
-rw-r----- 1 belling cdrom 4883517198 Feb 17 16:37 101208_I117_FC80AKJABXX_L3_HUMfrwTARAAPEI-5_1.fq
-rw-r----- 1 belling cdrom 4883517198 Feb 17 16:38 101208_I117_FC80AKJABXX_L3_HUMfrwTARAAPEI-5_2.fq
-rw-r----- 1 belling cdrom 5431922303 Feb 17 16:39 101208_I117_FC80AKJABXX_L3_HUMfrwTBRAAPEI-6_1.fq
-rw-r----- 1 belling cdrom 5431922303 Feb 17 16:39 101208_I117_FC80AKJABXX_L3_HUMfrwTBRAAPEI-6_2.fq
-rw-r----- 1 belling cdrom 5224871083 Feb 17 16:40 101208_I117_FC80AKJABXX_L3_HUMfrwTCRAAPEI-7_1.fq
-rw-r----- 1 belling cdrom 5224871083 Feb 17 16:40 101208_I117_FC80AKJABXX_L3_HUMfrwTCRAAPEI-7_2.fq
-rw-r----- 1 belling cdrom 5391641213 Feb 17 16:41 101208_I117_FC80AKJABXX_L3_HUMfrwTDRAAPEI-8_1.fq
-rw-r----- 1 belling cdrom 5391641213 Feb 17 16:41 101208_I117_FC80AKJABXX_L3_HUMfrwTDRAAPEI-8_2.fq
```

Quality Check

- Base quality
- Sequence length
- K-mers repeats
- ATGC ratio
- Occurrence of reads

Before trimming



Trimming

- Remove bad quality reads
- Parameters
  - Minimum length of reads
  - Quality threshold
    - Quality score = -10*log10(Error rate)

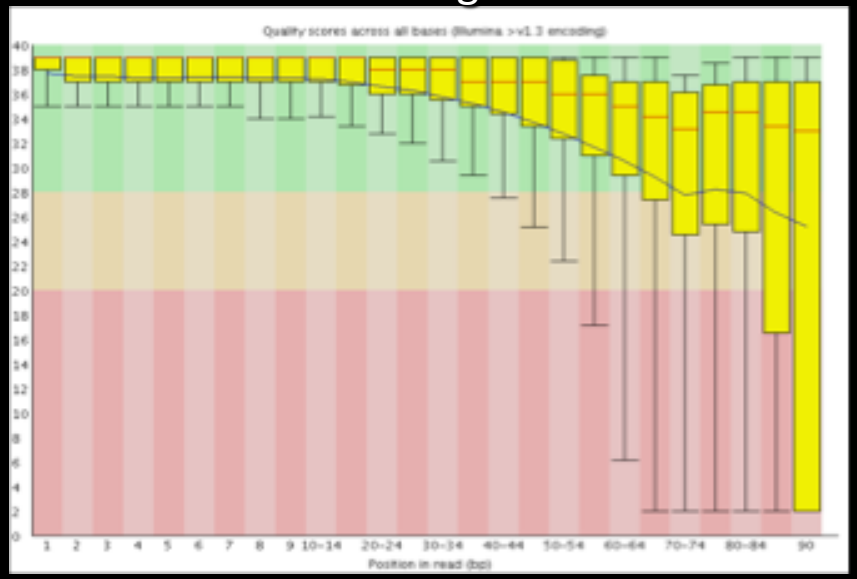
The Relationship Between Quality Score and Base Call Accuracy

Quality Score	Probability of Incorrect Base Call	Inferred Base Call Accuracy
10 (Q10)	1 in 10	90%
20 (Q20)	1 in 100	99%
30 (Q30)	1 in 1000	99.9%

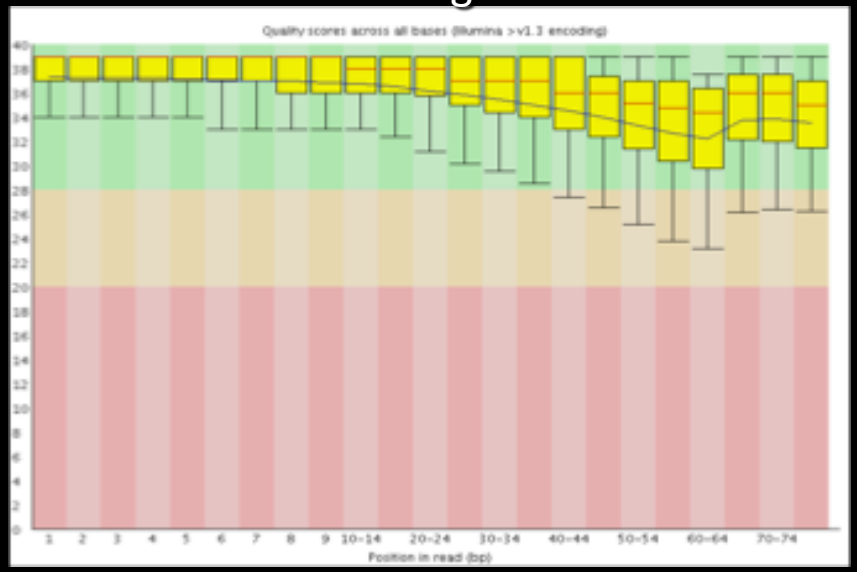
Quality Check

Per base sequence quality

Before trimming



After trimming

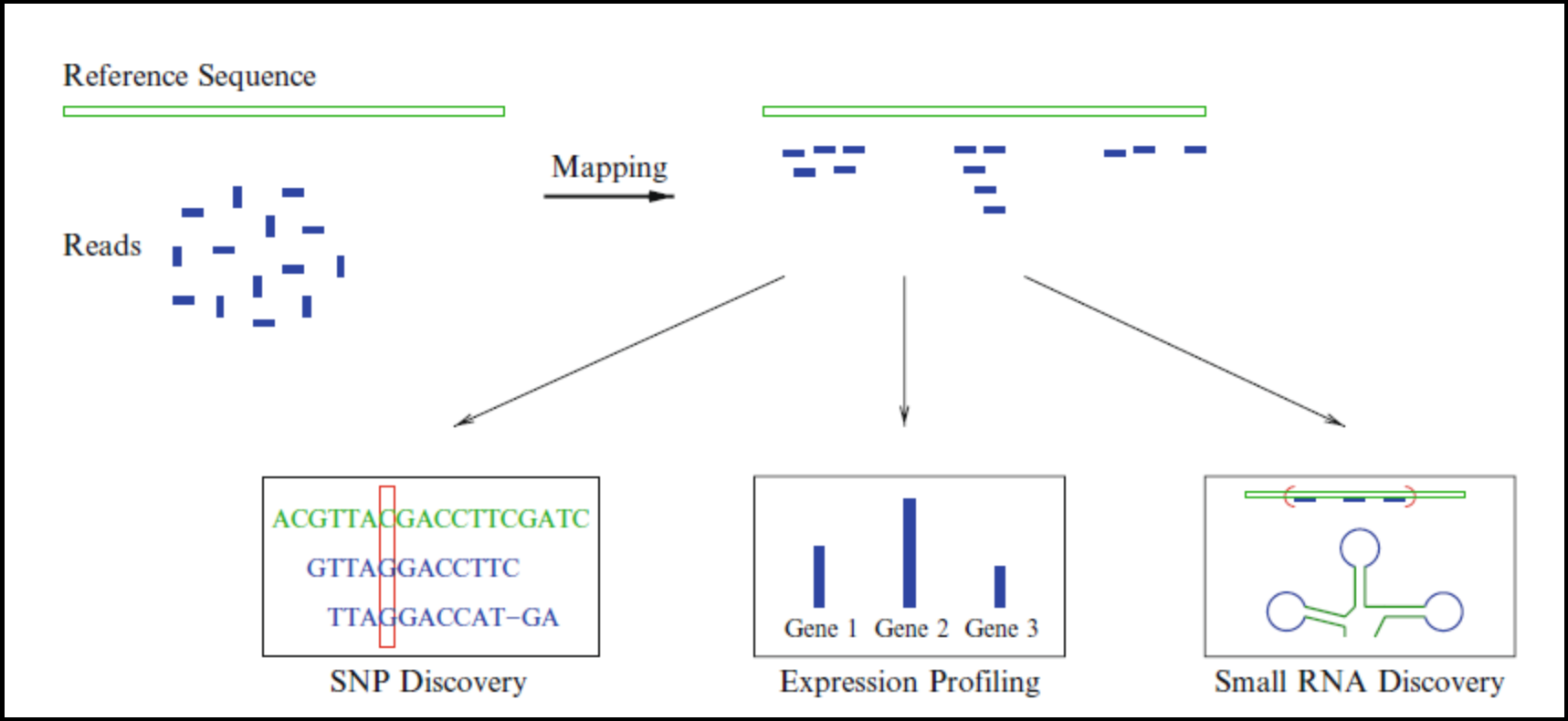


Alignment

- Reference based:
  - Reference genome
  - Annotations for the genome

- De-novo:
  - New genome (ex. Model organisms) and new discoveries.
  - Iterative process.

Reference mapping



Read mapping and its applications. Mapping programs are widely used to align reads to a reference while allowing some flexibility in terms of mismatches and indels and a policy for handling ambiguous matches

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                    ATAGCT..............TCAGC
-r003                          ttagctTAGGC
-r001/2                                     CAGCGCCAT
```

A

- ce Alignment/Map)

format is a generic
nucleotide sequenc
binary format of SA

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{29}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-1]	MAPping Quality
6	CIGAR	String	*|([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	*|=|[!-()+-<>-~][!-~]*	Ref. name of the mate/next segment
8	PNEXT	Int	[0,2^{29}-1]	Position of the mate/next segment
9	TLEN	Int	[-2^{29}+1,2^{29}-1]	observed Template LENgth
10	SEQ	String	*|[A-Za-z=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

Alignment terms

- Alignment score – A metric that represents how well a read maps to the reference genome (e.g. for a read of 100 bases, if it matches to the reference with 2 mismatches, the alignment score is 98).

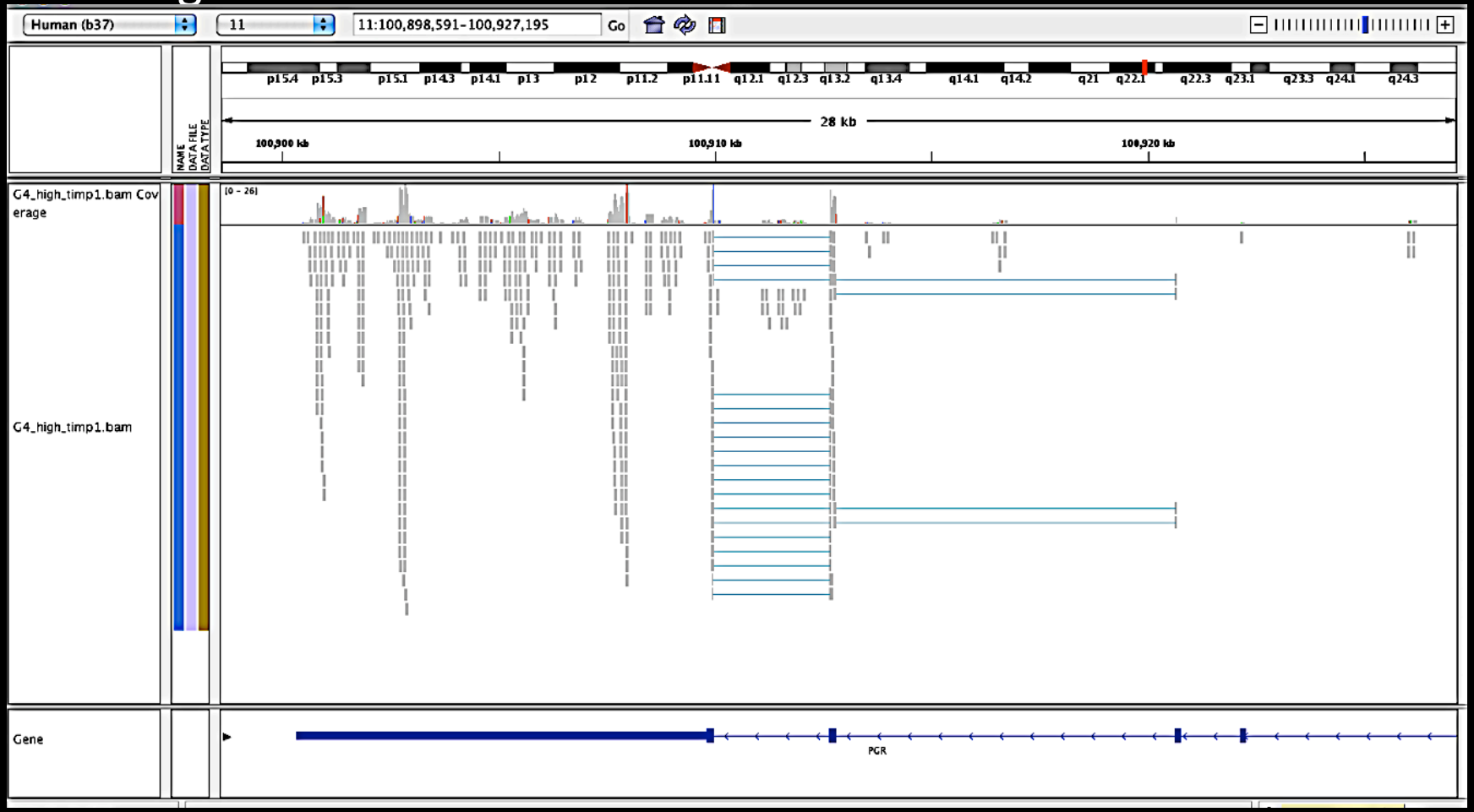
- CIGAR scores – A metric that indicates HOW a read maps to the reference (e.g. 98M2m means that a read of 100 bases matches to the reference with 98 matches [M] and 2 mismatches [m]).

- Mapping quality – A measure of the confidence that a read actually comes from the region it is aligned to by the alignment algorithm. A multiply mapping read has a low mapping quality and a uniquely mapping read has a high mapping quality.

- Multiply mapping reads – Reads that align to more than one location on the reference genome (either due to less stringent alignment criteria or due to repeats in the genome)

Alignment filtering/Visualization

Filtering for:



Alignment with reference genome

- Read depth – No. of times a particular base in the fragment of interest is read. An avg read dept of 20X means that in a particular dataset on an average a base has been sampled 20 times.

- Coverage – No. of bases of a genome that are sequenced (e.g. Out of the 3 bn human bases if one sequences 30 million bases then the coverage would be 30 Mb).

Note: Depth and Coverage are used interchangably!

Post-alignment analysis

Filter out bad aligned reads

Annotate SNPs/ Indels

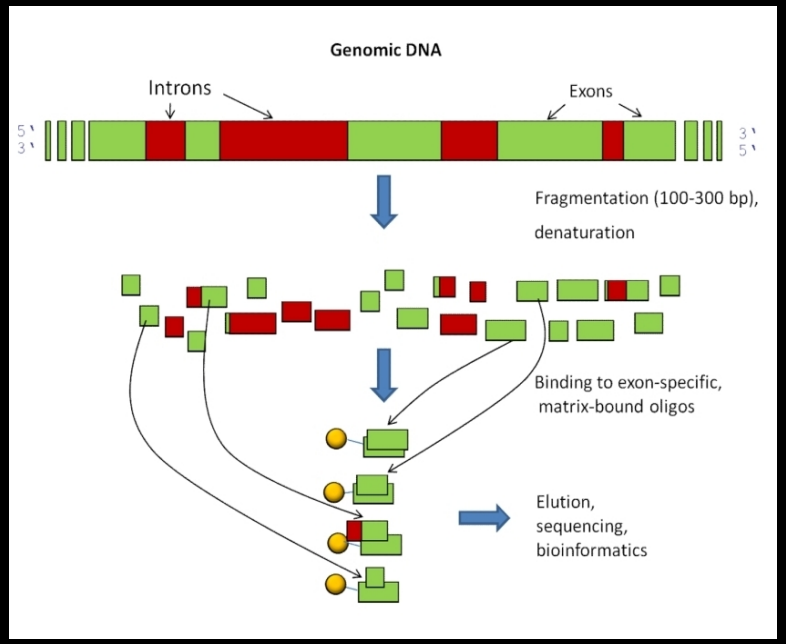
Gene expression estimates

- Comparative genomics
- Identify disease-causing variations

Introduction to exercise

Data

- Exome sequencing data
- Paired end reads
- Breast cancer cell line



BRCA2: *One of the most common germ line mutation*

A common variant in *BRCA2* is associated with both breast cancer risk and prenatal viability

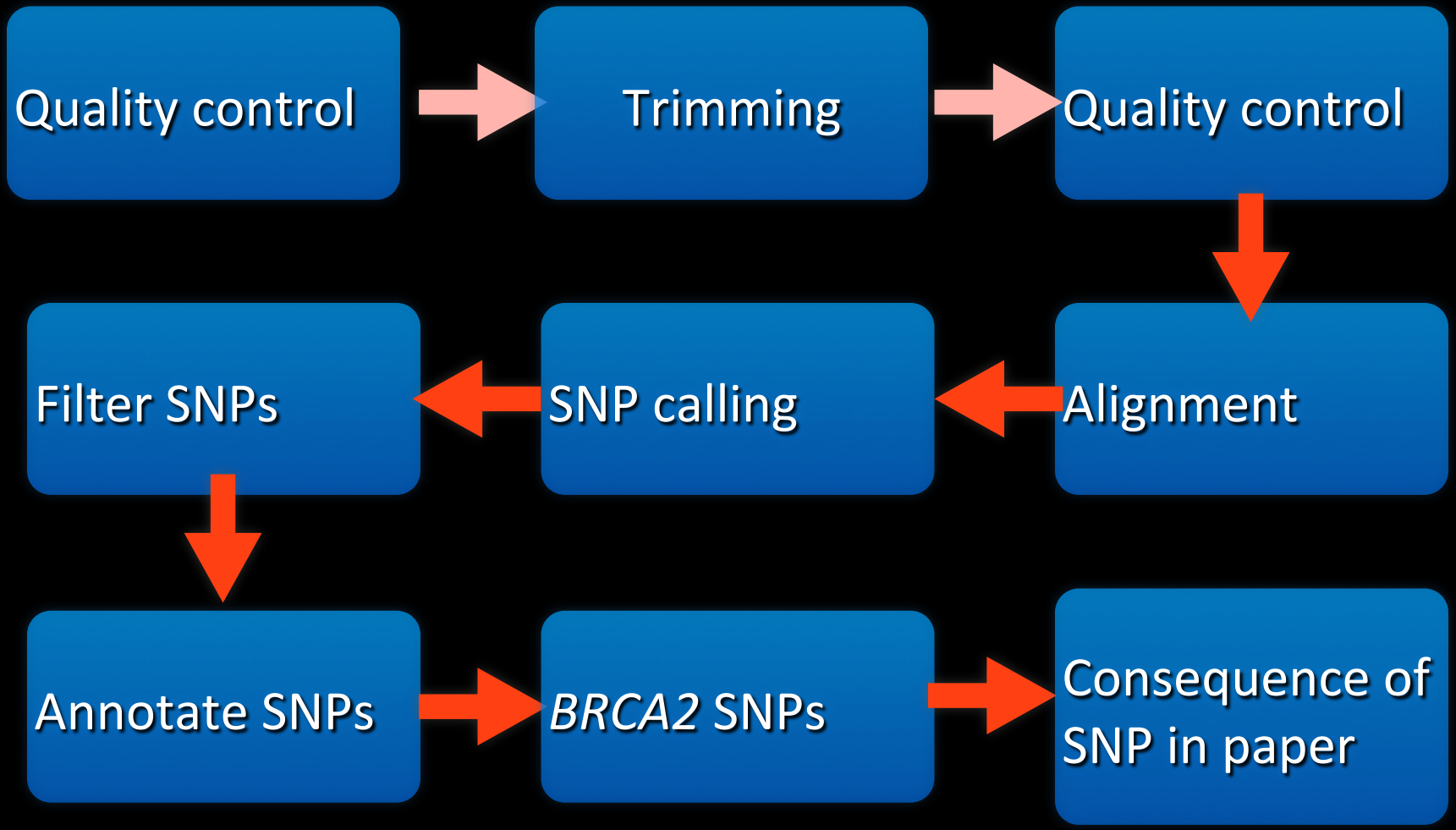
Catherine S. Healey[1]*, Alison M. Dunning[1]*, M. Dawn Teare[2], Diana Chase[4], Louise Parker[5], John Burn[6], Jenny Chang-Claude[7], Arto Mannermaa[8], Vesa Kataja[9], David G. Huntsman[10], Paul D.P. Pharoah[1], Robert N. Luben[3], Douglas F. Easton[2] & Bruce A.J. Ponder[1]

These authors contributed equally to this work.

Inherited mutations in the gene *BRCA2* predispose carriers to early onset breast cancer, but such mutations account for fewer than 2% of all cases in East Anglia. It is likely that low penetrance alleles explain the greater part of inherited susceptibility to breast cancer; polymorphic variants in strongly predisposing genes, such as *BRCA2*, are candidates for this role. BRCA2 is thought to be involved in DNA double strand break-repair[1,2]. Few mice in which *Brca2* is truncated survive to birth; of those that do, most are male, smaller than their normal littermates and have high cancer incidence[3,4]. Here we show that a common human polymorphism (N372H) in exon 10 of *BRCA2* confers an increased risk of breast cancer: the HH homozygotes have a 1.31-fold (95% CI, 1.07–1.61) greater risk than the NN group. Moreover, in normal female controls of all ages there is a significant deficiency of homozygotes compared with that expected from Hardy-Weinberg equilibrium, whereas in males there is an excess of homozygotes: the HH group has an estimated fitness of 0.82 in females and 1.38 in males. Therefore, this variant of BRCA2 appears also to affect fetal survival in a sex-dependent manner.

In an initial study to investigate whether common *BRCA2* variants alter the risk of breast cancer in the general population, we carried out an association study on six *BRCA2* polymorphisms identified through the BIC database (http://www.nhgri.nih.gov/Intramural_research/Lab_transfer/BIC; Table 1). The genotype distributions of both the exon 10 N372H and the 5′ UTR a-26g polymorphisms approached significant differences between cases and controls in our initial hypothesis-generating study. We thus restricted studies in further case-control series to confirmation of these observations, however, only the N372H findings were confirmed. N372H is the sole *BRCA2* variant resulting in an amino

Workflow of analysis



Galaxy

Job in queue

Job running

Job finished